CaFA: Cost-aware, Feasible Attacks With Database Constraints Against Neural Tabular Classifiers

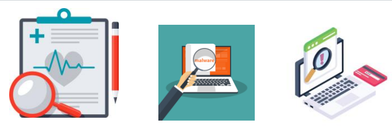
Matan Ben-Tov, Daniel Deutch, Nave Frost, Mahmood Sharif

Code

Paper

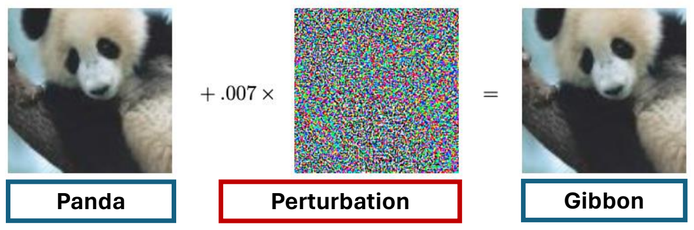
Motivation

ML on **tabular** data is ubiquitous in many **critical applications.**



Caveat: ML algorithms are susceptible to **evasion attacks**.

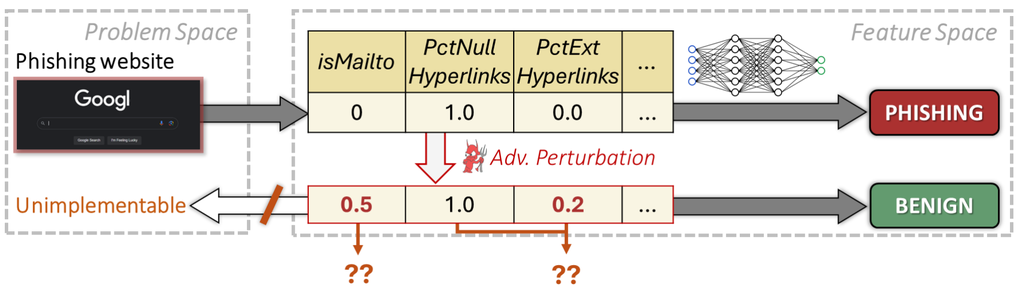
These attacks are thoroughly explored in **vision,** and can even be **realized** (as physical artifacts) generically.



Can we generically evaluate robustness against realizable evasion attacks in the tabular domain?

Naïve attempt, apply the same attack:

Main challenge – problem-feature space gap (heterogenous mapping that is neither differentiable nor invertible) [1]

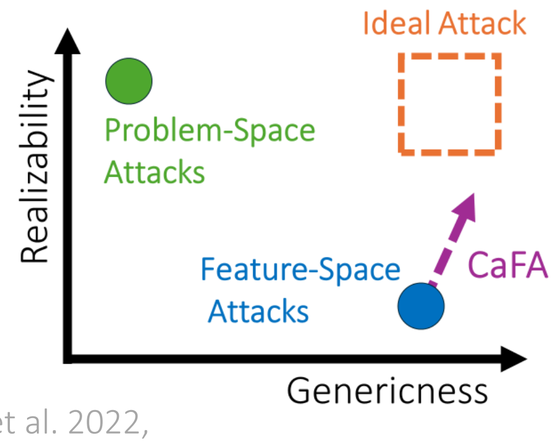


→ violated realizability and non-suitable cost. [1] Pierazzi et al. 2020

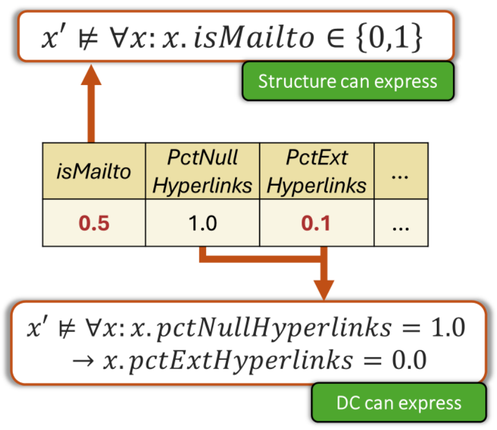
Background and Prior Work

Existing attacks (mainly) lack in realizability or genericness:

1. **Problem-Space Attacks**
   Manipulate problem-space instances *directly* via allowed transformations.
   E.g., Lucas et al. 2021, Eykholt et al. 2023
2. **Feature-Space Attacks**
   Manipulate feature-space instances + accounting for realizability and imperceptibility.
   E.g., Simonetto et al. 2022, Mathov et al. 2022, Sheatsley et al. 2021, Kireev et al. 2023



For **realizability**, we consider **data-integrity constraints types**:



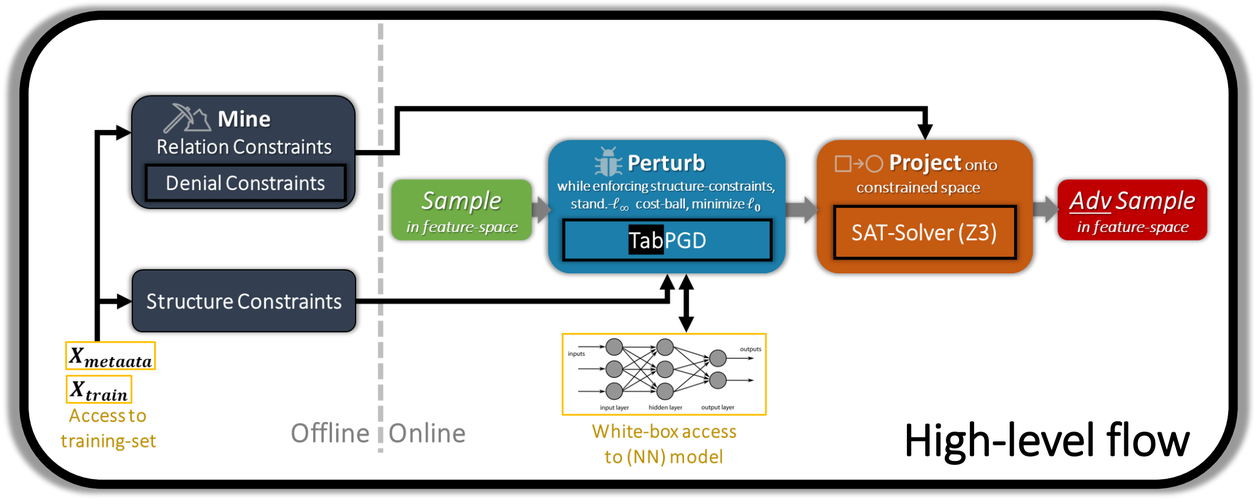
Structure Constraints
Involve structural properties
Feature's type, permissible range, etc.

Relational Constraints
Involve complex cross-feature relations
Denial Constraints (DCs):
- Widely used, highly expressive
- $\forall x, x' \in X: \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \neg(p_1 \wedge \cdots \wedge p_\ell)$
- Mined from the training-set as *soft* constraints

Our work offers a generic scheme for feature-space modifications that induces misclassification, is feasible in terms of feature-space constraints, and requires minimal adversarial effort.

Our Approach

CaFA is a three-stage generic system for cost-aware, feasible robustness evaluation of white-box neural tabular classifiers.



1. Mine T

Structure constraints	
isMailto	Categorical, {0,1}
PctExt Hyperlinks	Continuous, [0,1]
...	...

Mined using FastADC algorithm, allowing 1% violation rate

Relation constraints	
DC #1	$\neg(p_1 \wedge \cdots \wedge p_\ell)$
DC #2	...
...	...

- A **ranking-scheme** based on established metrics
- **Filtering-out** low-performing DCs

2. Perturb x

Heterogenous step size

Project on **structure constraints** and *standardized-*ℓ_∞ ϵ-ball

Freeze least *essential* feature **each run**

TabPGD:
i. $x'^{(0)} \coloneqq x$
ii. $g \coloneqq \nabla_x L(x'^{(i)}, y)$
iii. $\delta \coloneqq \overleftarrow{R}\alpha \cdot sign(g)$
iv. $x'^{(i+1)} \coloneqq x'^{(i)} + \Pi(\delta)$

TabCW-L0:
$p_j(\delta, g, R) = \left|\frac{g_j \cdot \delta_j}{R_i}\right|$

3. Project x' onto T

Search done with **Z3 SAT Solver**

Binary search to minimize # (least constrained) literals to relax

Project $if: x' \notin T$
- Relax literals in x'
- Find values such that $x' \models T \wedge cost$
- **Minimize** relaxed literals

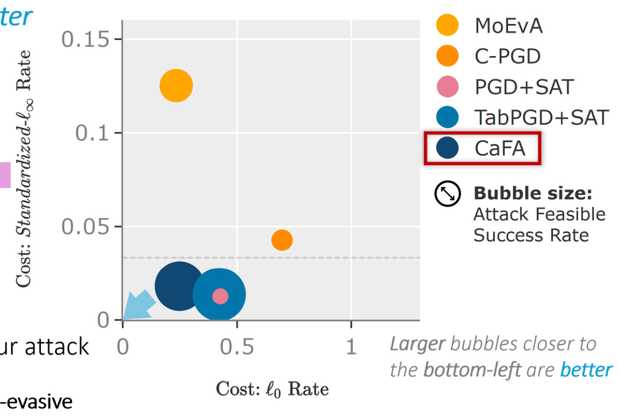
Results

Evaluated on *adult*, *bank*, *phishing* datasets and MLPs and TabNets.

Constraints Used	Completeness	Soundness	F1
Denial Constraints (Our work)	88.3%	75.1%	81.2%
Valiant's (Sheatsley et al.)	99.1%	7.1%	14.7%

↑ *higher are better*

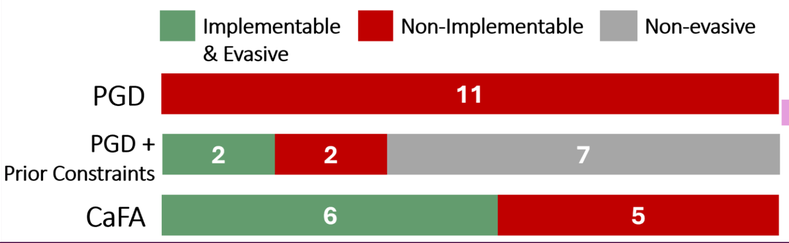
Denial Constraints are **high-quality constraints**



CaFA has **lowest combined-cost and highest feasible success**

Larger bubbles closer to the bottom-left are better

Attempted to implement **real-world phishing** evasion using our attack



CaFA can be **realized**